

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-132453

(43)Date of publication of application : 10.05.2002

(51)Int.Cl.

G06F 3/06
G06F 12/08

(21)Application number : 2000-321280

(71)Applicant : HITACHI LTD

(22)Date of filing : 20.10.2000

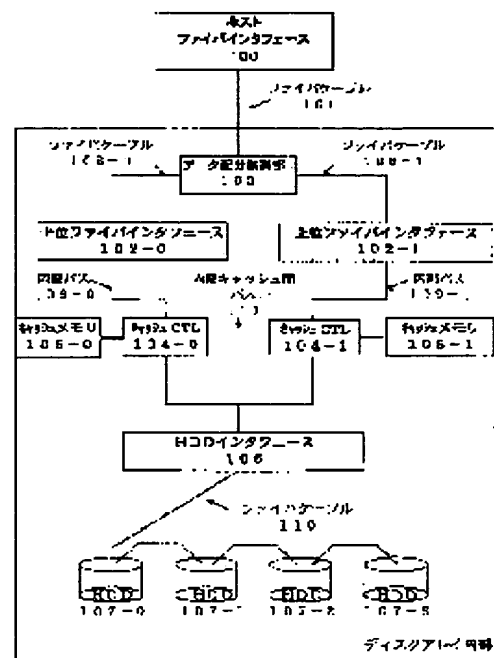
(72)Inventor : KYO SHIYUMEI
YAMANASHI AKIRA
YAGISAWA IKUYA

(54) DISK ARRAY SYSTEM

(57)Abstract:

PROBLEM TO BE SOLVED: To realize a disk array having a high internal transfer capacity corresponding to a host computer with a high-speed fiber interface.

SOLUTION: This disk array system is equipped with plural upper fiber interfaces 102 that receive/send data from/to a host fiber interface 100, a cache memory 105 connected to the plural upper fiber interfaces, plural disk storage means 107 which data is written on or read from, and a disk drive interface 106 for controlling the storage means, and each of the plural upper fiber interfaces has a cache memory. Moreover, in the disk array system, a data distribution controlling part 103 is provided between the host fiber interface and the upper fiber interfaces.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

(19)日本国特許庁 (J P)

(12) 公 開 特 許 公 報 (A)

(11)特許出願公開番号
特開2002-132453
(P2002-132453A)

(43)公開日 平成14年5月10日(2002.5.10)

| (51)Int.Cl. ⁷ | 識別記号 | F I | テマコード*(参考) |
|------------------------------|-------------------------|--------------|--|
| G 0 6 F 3/06 | 3 0 1 | G 0 6 F 3/06 | 3 0 1 U 5 B 0 0 5 3 0 1 T 5 B 0 6 5 |
| 12/08 | 5 4 0 5 1 1 5 5 7 | 12/08 | 5 4 0 5 1 1 Z 5 5 7 |
| 審査請求 未請求 請求項の数8 O L (全 12 頁) | | | |

(21)出願番号 特願2000-321280(P2000-321280)

(22)出願日 平成12年10月20日(2000.10.20)

(71)出願人 000005108

株式会社日立製作所
東京都千代田区神田駿河台四丁目6番地

(72)発明者 姜 小明

神奈川県小田原市国府津2880番地 株式会
社日立製作所ストレージシステム事業部内

(72)発明者 山梨 晃

神奈川県小田原市国府津2880番地 株式会
社日立製作所ストレージシステム事業部内

(74)代理人 100093492

弁理士 鈴木 市郎 (外1名)

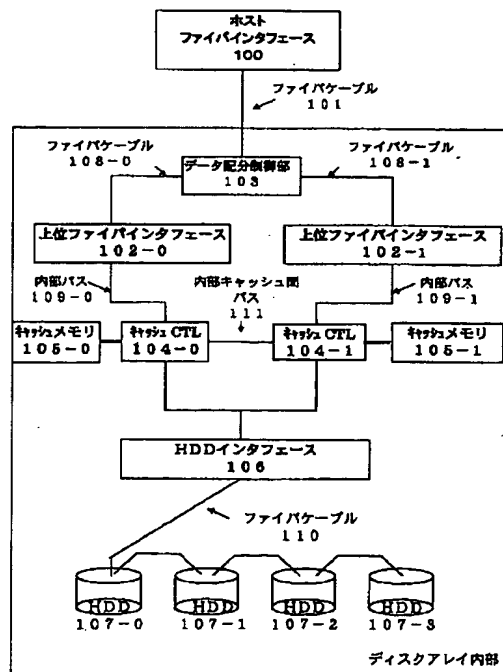
最終頁に続く

(54)【発明の名称】 ディスクアレイシステム

(57)【要約】

【課題】 高速ファイバインタフェースを持つホストコンピュータに対応した高い内部転送能力を有するディスクアレイシステムを実現すること。

【解決手段】 ホストファイバインタフェース100からのデータを授受する複数の上位ファイバインタフェース102と、前記複数の上位ファイバインタフェースに接続されたキャッシュメモリ105と、データを書き込み又は読み出す複数のディスク記憶手段107と、前記記憶手段を制御するディスクドライバインタフェース106と、を備えたディスクアレイシステムであって、前記複数の上位ファイバインタフェースの各上位ファイバインタフェースに対応してそれぞれキャッシュメモリを備えている。また、前記ディスクアレイシステムにおいて、ホストファイバインタフェースと上位ファイバインタフェースとの間にデータ配分制御部103を設ける。



【特許請求の範囲】

【請求項 1】 ホストファイバインタフェースからのデータを授受する複数の上位ファイバインタフェースと、前記複数の上位ファイバインタフェースに接続されたキャッシュメモリと、データを書き込み又は読み出す複数のディスク記憶手段と、前記記憶手段を制御するディスクドライバインタフェースと、を備えたディスクアレイシステムであって、

前記複数の上位ファイバインタフェースの各上位ファイバインタフェースに対応してそれぞれキャッシュメモリを備えていることを特徴とするディスクアレイシステム。

【請求項 2】 請求項 1 に記載のディスクアレイシステムにおいて、

前記ホストファイバインタフェースと前記上位ファイバインタフェースとの間に設けられたデータ配分制御部は複数のデータバッファを備え、

前記データ配分制御部は、前記キャッシュメモリの空き容量の管理を行い、前記ホストファイバインタフェースからのデータを前記データバッファへ格納し、又は前記データバッファのバッファデータを前記キャッシュメモリに転送する、機能を有することを特徴とするディスクアレイシステム。

【請求項 3】 請求項 2 に記載のディスクアレイシステムにおいて、前記データ配分制御部はキャッシュメモリ内のキャッシュ管理テーブルのコピーを有することを特徴とするディスクアレイシステム。

【請求項 4】 請求項 1 に記載のディスクアレイシステムにおいて、前記各上位ファイバインタフェースに対応してそれぞれ備えられたキャッシュメモリ同士間でデータ及び管理情報を授受するための通信バスを有することを特徴とするディスクアレイシステム。

【請求項 5】 ホストファイバインタフェースからのデータを授受する複数の上位ファイバインタフェースと、前記複数の上位ファイバインタフェースに接続されたキャッシュメモリと、データを書き込み又は読み出す複数のディスク記憶手段と、前記記憶手段を制御するディスクドライバインタフェースと、を備えたディスクアレイシステムであって、

前記ホストファイバインタフェースと前記上位ファイバインタフェースとの間にデータ配分制御部を設け、前記データ配分制御部は、前記ホストファイバインタフェースからのデータを複数の分割して前記複数の上位ファイバインタフェースに供給するとともに、前記複数の上位ファイバインタフェースを介して前記分割されたデータを並列に複数のキャッシュメモリに同時に転送することを特徴とするディスクアレイシステム。

【請求項 6】 請求項 5 に記載のディスクアレイシステム

ムにおいて、

前記データ配分制御部は複数のデータバッファを備え、前記データ配分制御部は、前記キャッシュメモリの空き容量の管理を行い、前記ホストファイバインタフェースからのデータを前記データバッファへ格納し、又は前記データバッファのバッファデータを前記キャッシュメモリに転送する、機能を有することを特徴とするディスクアレイシステム。

【請求項 7】 請求項 5 に記載のディスクアレイシステムにおいて、

前記データ配分制御部はキャッシュメモリ内のキャッシュ管理テーブルのコピーを有することを特徴とするディスクアレイシステム。

【請求項 8】 請求項 5 に記載のディスクアレイシステムにおいて、

前記各上位ファイバインタフェースに対応してそれぞれ設置されたキャッシュメモリ同士間でデータと管理情報を授受するための通信バスを有することを特徴とするディスクアレイシステム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明はファイバインタフェースを使用するコンピュータシステムに接続される磁気ディスク装置、磁気テープ装置、半導体記憶装置または光ディスク装置などの記憶媒体を制御する記憶システムに関する。

【0002】

【従来の技術】ディスクアレイシステムは、RAID (Redundant Array of Inexpensive Disks) と呼ばれ、複数のディスク装置を列状 (アレイ状) に配置した構成をとり、ホストからのリード要求 (データの読み出し要求) およびライト要求 (データの書き込み要求) をディスクの並列動作によって高速に処理するとともに、データに冗長データを付加することによって信頼性を向上させた記憶装置である。ディスクアレイシステムは、冗長データの種類と構成により 5 つのレベルに分類されている (論文: 「A Case for Redundant Arrays of Inexpensive Disks (RAID)」, David A. Patterson, Garth Gibson, and Randy H. Katz, Computer Science Division Department of Electrical Engineering and Computer Sciences, University of California Berkeley)。

【0003】ホストからのライト要求時には、ディスクアレイ内部では、ホストから受け取ったデータを特殊なアドレス変換によって分割し、一時的にキャッシュに格

10

20

30

40

50

納後、順次、ディスク装置へと並列に書込んでいく方式が一般的である。逆に、ホストからのリード要求時には、分割したデータをディスク装置から並列に読み出してキャッシュへと格納し、さらにキャッシュから読み出した分割データを元のデータとして束ね、ホストへと転送する。

【0004】図2はRAID技術による従来ディスクアレイ構成図である。ディスクアレイシステムは、ホストインタフェース200に接続するディスクアレイのファイバまたはSCSIの上位インタフェース202と、イン

タフェース202と内部バス203で互いに接続されるMPUユニット204及びキャッシュコントローラ205と、キャッシュコントローラ205に接続するHDDインタフェース207と、HDD群209と、から構成される。

【0005】図2では、上位インタフェースを1個として書いているが、複数であってもよい。この場合に、図2の構成において、共有するキャッシュメモリに複数の上位インタフェースが接続される構成となっており、上位インタフェースの転送能力を活かしきるため、上位イン

タフェースとキャッシュメモリ間のデータバスの転送能力が上位インタフェース転送能力以上であることが前提となる。ここで言う転送能力とは、単位時間あたりに送ることのできるデータ量のことであり、例えば、1秒間に1ギガバイトのデータを転送できる能力を1ギガバイト毎秒（GB/s）と表すこととする。

【0006】

【発明が解決しようとする課題】図2に示す従来アーキテクチャでは、ホストからのリード要求において、ディスクからキャッシュメモリにデータを送り、その後、キャッシュメモリからホストにデータを送るという流れである。ホストインタフェースの転送能力を1とした場合、ディスクインタフェースの転送能力は1、キャッシュメモリの転送能力は2となる。キャッシュメモリは、データの入力と出力があるため、ホストインタフェース、および、ディスクインタフェースの2倍の転送能力となる。ホストインタフェースの転送能力を生かしきるためには、キャッシュメモリの転送能力はホストインタフェースの2倍が必要となる。同様に、上位ホストの個数がN倍となれば、上位ホストインタフェースの転送能力もN倍となり、キャッシュメモリの転送能力は2N倍となる。

【0007】このことから分かるように、従来ディスクアレイのアーキテクチャにおいて、上位側の転送能力以上に、内部バスの転送能力を上げると共に、キャッシュメモリの転送能力も内部バス転送能力の向上に伴って上げるのが必須である。

【0008】本発明の目的は、前述した従来技術の課題を解決し、複数の上位ファイバインタフェースを同時に駆使し、一つの高速度ホストファイバインタフェースとデ

ータやり取りを行うことにより、ホストファイバインタフェースに比べてその転送速度の低い内部バスを有するディスクアレイでも、転送速度が高速度であるホストファイバインタフェースへの対応を可能とすることにある。

【0009】また、本発明の他の目的は、ディスクアレイ内部で複数上位ファイバインタフェースを有する場合、ホストに複数ファイバインタフェースをアクセスする時の管理、或はターゲットID（複数ファイバのそれぞれのポートに付したアドレス）の切替負荷を掛けずに（ターゲットIDを切り替えた後にファイバ接続するのではホストに負荷が掛かるので）、一つのファイバインタフェースをアクセスするようにする手段を提供することにある。

【0010】

【課題を解決するための手段】前記課題を解決するために、本発明は主として次のような構成を採用する。

【0011】ホストファイバインタフェースからのデータを授受する複数の上位ファイバインタフェースと、前記複数上位ファイバインタフェースに接続されたキャッシュメモリと、データを書き込み又は読み出す複数のディスク記憶手段と、前記記憶手段を制御するディスクドライバインタフェースと、を備えたディスクアレイシステムであって、前記複数の上位ファイバインタフェースの各上位ファイバインタフェースに対応してそれぞれキャッシュメモリを備えているディスクアレイシステム。

【0012】また、ホストファイバインタフェースからのデータを授受する複数の上位ファイバインタフェースと、前記複数上位ファイバインタフェースに接続されたキャッシュメモリと、データを書き込み又は読み出す複数のディスク記憶手段と、前記記憶手段を制御するディスクドライバインタフェースと、を備えたディスクアレイシステムであって、前記ホストファイバインタフェースと前記上位ファイバインタフェースとの間にデータ配分制御部を設け、前記データ配分制御部は、ライト時には、前記ホストファイバインタフェースからのデータを複数に分割して前記複数の上位ファイバインタフェースに供給するとともに、前記複数上位ファイバインタフェースを介して前記分割されたデータを並列に複数キャッシュメモリに同時に転送する、また、リード時においては、複数キャッシュメモリからのデータを並列に複数の上位ファイバインタフェースを通してデータ配分制御部に転送して前記データ配分制御部で分割されたデータを元のデータとして束ね、ホストへリードデータを返すディスクアレイシステム。

【0013】

【発明の実施の形態】本発明の実施形態に係るディスクアレイシステムについて、図面を用いて以下説明する。図1は本発明の実施形態に係るディスクアレイシステムであり、ディスクアレイシステムは、図1に示すように、ホストファイバインタフェース100にファイバケ

ープル101を介して接続されるデータ配分制御部103と、前記データ配分制御部103にファイバケーブル108-0、1を介して接続される上位ファイバインタフェース102-0と102-1と、上位ファイバインタフェース102-0、1に接続される一時的にデータを保存するキャッシュメモリの制御コントローラ104-0と104-1と、これらのキャッシュコントローラ104-0、1配下のキャッシュメモリ105-0、1と、HDDインタフェース106と、ディスク群HDD107と、から構成される。キャッシュコントローラ104-0と104-1間でキャッシュ管理情報とデータの通信を出来るように、高速なデータバス111により接続される。

【0014】前述したディスクアレイの構成において、ホストは、転送速度がディスクアレイのそれよりも高速であるファイバインタフェースを介してディスクアレイに対してライトする時、ディスクアレイのデータ配分制御部103で一旦データを受け取り、ディスクアレイ内部の上位ファイバインタフェースの数に応じてデータを分割する。また、受け取ったホストコマンドから上位ファイバインタフェースへのライトコマンドを生成し、各上位ファイバインタフェース102-0、1を介してキャッシュメモリ105-0、1にライトする。

【0015】キャッシュメモリ105で分割されたデータブロックにホストデータを復元するための関連性を持たせた後、冗長情報をつけ、空き時間でHDD群107にライトする。また、ホストからのリード命令に対しては、キャッシュメモリ上にデータが存在する場合はキャッシュメモリ105から、キャッシュメモリにデータが存在しない時は、HDD群107より、データ配分制御部103に要求データを送り、データ配分制御部103で連続したデータにし、ホストに返す（詳細は後述）。

【0016】次に、図3に基づいてデータ配分制御部103の機能乃至作用を説明する。図3に示す構成は、図1のデータ配分制御部103の詳細構造である。ホストファイバインタフェース100と上位ファイバインタフェース102間の転送スピード差を十分に吸収するため、各上位ファイバインタフェースにおいて、2面のバッファを設け、総計4面とする。各バッファの容量は等しい。ホストファイバドライブ制御部301に4面のデータ転送バッファ304-0、1、2、3が、301のホストファイバと同速度を有する内部バス303、305-0、1、2、3とバッファバススイッチ302を介して接続され、また、前記の4面データバッファが上位ファイバドライブ制御部307と同等な転送速度を有する内部データバス308、309、311-0、1、2、3とバススイッチ306を介して上位のファイバドライブ制御部307-0、1に接続される。

【0017】バッファスイッチ302によりホストファイバドライブ制御部301が接続したいバッファへの切

り替えを行い、301が使用していないほかの3面バッファを上位ファイバドライブ制御部307-0、1に使用させることを可能とする。また、バススイッチ306の切り替えにより、4面バッファの内の2面が同時にファイバドライブ制御部307-0、1との間でデータ転送を行うことができる。前記バススイッチの切り替え動作は後述の転送リストに基づいて行われる。バススイッチを導入することにより、バス転送能力は勿論、バッファとファイバドライブ制御部間のデータ転送速度が最大限に引き出せる。

【0018】MPUユニット部310にホストコンピュータからのコマンドを記憶するエリアを設ける。アクセス先のアドレス及び転送データ長をホストから受け取ったコマンドより割り出し、転送データ長を上位ファイバインタフェースの数で割った値が各上位ファイバドライブ制御部307-0、1が転送するデータの長さとなり、更にこの転送データ長を用いてデータ配分制御部内の一つバッファの大きさを割ったものを各上位ファイバドライブ制御部307-0、1が起動するバッファの回数とする。例えば、上位ファイバインタフェースの数はN、バッファのサイズはMとして、ホストからのリードまたはライトデータの長さはLとする場合は、各上位ファイバインタフェースによる転送データ長は L/N となり、各上位ファイバインタフェースが起動するバッファの回数は $(L/N)/M$ となる。前記転送回数を用いてホストファイバドライブ制御部301、上位ファイバドライブ制御部307に転送リストを生成する。

【0019】図8は上位ファイバドライブ制御部307における転送リストの例であり、上位ファイバドライブ307の制御部はこの転送リストに従ってバッファからキャッシュへデータ転送を行う。図8の(1)はファイバドライブ制御部307-0における制御部転送リストであり、図8の(2)はファイバドライブ制御部307-1における制御部転送リストである。

【0020】また、2面のキャッシュメモリの管理テーブルのコピーをMPUユニット310に設け、キャッシュメモリにある管理テーブルが更新された場合は、その更新が上位ファイバインタフェースを通じてMPUユニット310にある管理テーブルに反映される。ホストからのライトコマンドに対し、バッファの容量ではなく、キャッシュ管理テーブルを検索し、キャッシュメモリの空き容量を見せることにより、ホストからの大容量なデータ転送要求を高速に応じることが可能となる。

【0021】内部バッファの管理を単純化するために、バッファに連続したアドレスを持たせる。ホストファイバドライブ制御部301と上位ファイバドライブ307の間でバッファの使用衝突を回避するため、バッファ使用中で有るかどうかのフラグビットを設ける。バッファ使用中であれば、該当バッファのフラグビットを0（使用不可の印）にし、使用完了後（使用バッファ中のデー

タの出力後)、該当バッファのフラグビットを1(使用可の印)にし、ファイバドライブがバッファ使用可の状態とする。たとえば、ホストファイバドライブ301からバッファ1(304-1)にデータをライトする時はバッファ1のフラグが0となり、上位ファイバドライブ制御部307からはバッファ1へのアクセスができないが、バッファ2、304-2のフラグが1であれば、上位ファイバドライブからのリードまたはライトは可能である。

【0022】ホストからのライト処理の流れを図6を用いて説明する。ここで、バッファ604-0、1、2、3の容量を300KBと仮定する。ホストからのライト要求は1.8MBで、バッファ総容量より大きい場合の流れを示す。

【0023】データ配分制御部にてホストからのライトコマンドを解釈し、ライト先アドレスLBA(Logical Block Address)と転送データ長LengthをMPUユニット部609のメモリに一時保存する。同時にキャッシュの容量を確認して空き容量が要求データ長より大きい場合に、ホストに全数転送の応答を返し、一回で1.8MBのデータを受け取る。キャッシュ容量が少なく、1.8MBデータを一回で受け取れない場合はホストにキャッシュ空き容量分データ転送の要求を返す。また、キャッシュに空きができた時点で、残りのデータを受信する。図示する例では、典型的なキャッシュ空き容量が十分あるケースのみ紹介する。

【0024】ホストからの1.8MBのデータをホストファイバドライブ制御部601よりバッファ0に300KB、バッファ1に300KB、バッファ2に300KB、バッファ3に300KB、バッファ0に300KB、バッファ1に300KBを格納していく。バッファ切替のタイムチャットを図4に示す。

【0025】ホストファイバドライブ制御部601がバッファ0(604-0)へのデータ格納が完了したら、前記バッファ0のフラグを1にし、バッファ1(604-1)へのデータ格納を始める。それと同時に上位ファイバドライブ0(608-0)にホストからのライトアドレスLBAと上位ファイバドライブ608による転送データ長Length/2情報を含むライトコマンドを発行し、上位ファイバドライブ0(608-0)よりバッファ0(604-0)にあるデータをキャッシュ0(612-0)へ転送し始め(フラグを0にして)、転送完了後、バッファ0(304-0)のフラグビットを1に戻し、他のファイバドライブ制御部が使用できる状態にする。

【0026】ホストファイバドライブ制御部601がバッファ1(604-1)へのデータ格納終了後、バッファ2(604-2)へのデータ格納を続けるが、それと同時に、前記と同様に、上位ファイバドライブ1(608-1)に608-0に発行したライトコマンドを発行

し、ファイバドライブ1(608-1)よりバッファ1(604-1)のデータをキャッシュ1へ転送を行う。ホストファイバドライブ制御部601がバッファ2(604-2)へのデータ格納終了後、バッファ3(304-3)へのデータ格納を続け、それと同時に、上位ファイバドライブ0(608-0)がバッファ2(604-2)のデータを引き続きキャッシュ0へ転送する。キャッシュ0へ転送終了の後、バッファ2(604-2)のフラグビットを1にセットする。並行してホストファイバドライブ制御部より、バッファ0(604-0)へホストデータを入れ始める。ホストファイバドライブ制御部はバッファ3(604-3)へのデータ格納が終了後は、バッファ3のデータがファイバドライブ1(608-1)よりキャッシュ1へ転送される。

【0027】ホストファイバドライブ制御部601が最後の300KBをバッファ1に格納すると同時に上位ファイバドライブ0(608-0)がバッファ0(604-0)のデータをキャッシュへ転送する。ホストファイバドライブがバッファ1へのデータ格納は完了後は、上位ファイバドライブ1(608-1)より、バッファ1のデータをキャッシュ1へ格納するという繰り返しになる。

【0028】上記のように複数面バッファを使いまわすことにより、ホストからのデータ転送を中断せずに、ホストデータを900KBの二つのブロックに分割してキャッシュに格納することが可能となった。キャッシュがライト終了後に各上位インタフェースにライト完了の報告を上げ、データ配分制御部で各上位インタフェースからライト終了の信号が来たらホストにライト完了を報告する。

【0029】上記方法の実現は、上位ファイバドライブ608のトータルな転送能力がホストファイバドライブ601の転送能力より大きいことにより保証される。ホストとデータ配分制御部間(図5の(1))、データ配分制御部と上位ファイバインタフェース間(図5の(2))のデータ転送タイムチャットを図5に示す。図5の(1)に示す上下平行線の左側にはホストファイバインタフェースにおける送信又は受信内容が経時的に上から順に記載されており、その右側にはデータ配分制御部における受信又は送信内容が同様に記載されている。図5の(2)にはデータ配分制御部と上位ファイバインタフェース部との送信又は受信内容が図5の(1)と同様に経時的に記載されている。

【0030】図6において上位ファイバドライブ0'制御部(610-0)配下のキャッシュ0(612-0)に分割されたデータブロック0、900KBが格納され、ファイバドライブ1'制御部(610-1)配下のキャッシュメモリ1(612-1)にはブロック1、900KBが格納される。ここで、図6に示す上位ファイバドライブ0'制御部610-0及び上位ファイバドラ

イブ1' 制御部610-1は、図1に示す上位ファイバインタフェース102-0及び上位ファイバインタフェース102-1に対応するものである。

【0031】キャッシュの管理を簡単化するため、2面のキャッシュをアドレス連続な一面のキャッシュメモリとして管理することにより、現在のキャッシュ制御アルゴリズムはそのまま流用可能である。キャッシュメモリ0, 1にそれぞれキャッシュ管理テーブルSGCB0, 1 (Segment Control Block) を設け、キャッシュ0のテーブルをマスタとし、かつ、キャッシュ0にキャッシュ1の管理テーブルのコピーを持たせる。キャッシュ上での管理単位はセグメントであり、ブロック0, 1に2分割された1.8MBのホストデータはさらに、セグメント単位でキャッシュ上で分割され、SGCB0, 1より管理される。セグメントはキャッシュ上にある配置は連続でなくてもよい。SGCBにはセグメントのアドレスを持っているほか、キャッシュにステージしたデータのLBA (Logic Block Address) などの情報を有する。

【0032】キャッシュにデータが書かれた後はSGCB1の変化分をキャッシュ間のバスを介してキャッシュ0にあるSGCB1のコピーに反映される。マスタSGCB0が同一LBAを有するSGCB1に属するセグメントを、ホストライトデータの内の一部とみなし、SGCBは図7のように再構築される。よって、ホストデータが統合して管理される。図7に示す符号において、SEGLのSEGはセグメント (Segment) の意味であり、SEGLのLは左 (Left) の意味である。キャッシュが2面あって位置的に左側のキャッシュ0にあるSEGがSEGLとし、右にあるキャッシュ1のSEGがSEGRである。また、SEGL又はSEGRに続くADRはアドレスの意味である。

【0033】図7に示すように、ホストからのデータが左又は右キャッシュに分割されるが、分割されたデータが各キャッシュでSEG単位で管理される。2面キャッシュのSEGに図7に示す関係を持たせると、分散したデータが分散しているキャッシュで1つのデータとして管理、使用することが可能となる。

【0034】キャッシュコントローラ0, 1 (612-0, 1) にはデータ冗長性を付けする機能を持たせ、規格のRAID3又はRAID5の時はHDD群に保存のための冗長性コードをキャッシュ内で生成するときには、キャッシュ0, 1 (612-0, 1) が同期を取りながら、冗長コードを生成する。キャッシュ0 (612-0) でブロック0のパリティを算出後、そのパリティをと必要なデータをキャッシュ1 (612-1) に渡して、前記キャッシュ1でブロック1のデータパリティを合成する。HDDインタフェース (図1の106) が冗長性のあるデータブロック0, 1をHDD群 (図1の107) への書き込みを担当する。

【0035】データブロック0をHDD群 (107) へ書きこみ先はSGCB0が持っているLBAとデータ転送長情報より取り決めれる。ブロック1のHDD群への書きこみ先はSGCB1が持っているLBA+冗長性のあるブロック0の長さとしたLBAより取り決められる。HDD群へデータをライトする際は、先にブロック0、次にブロック1との順で行われる。RAID1の時は、キャッシュ間でデータコピーを行い、両キャッシュにブロック0, 1のある状態に達した後、HDDインタフェースより各配下のHDDへの書きこみを行う。

【0036】次に、ホストからの書きこみ要求が200KBであって、バッファ総容量より小さい場合の流れについて述べる。ホストからの200KBのデータをホストファイバドライバ制御部601よりバッファ0 (604-0) に100KB、バッファ1 (604-1) に100KBずつ格納し、バッファ2, 3を使用せず。ホストファイバドライバ制御部601がバッファ0 (604-0) へのデータ格納を完了後、バッファ0のフラグを1にし、バッファ0 (604-0) を他ファイバドライバ制御部が使用可の状態に設置する。

【0037】その後バッファ1 (604-1) へのデータ格納の開始と同時に上位ファイバドライバ0 (608-0) が前記バッファ0のデータを指定されたキャッシュ0 (612-0) に転送し始め、転送完了後、バッファ0のフラグビットを1に戻し、他ファイバ制御部使用できる状態にする。ホストファイバドライバ制御部601がバッファ1 (604-1) へのデータ格納が終了後、バッファ1を使用可の状態にした後、同様に、上位ファイバドライバ1が前記バッファ1のデータを指定されたキャッシュ1 (612-1) にデータ転送を行う。キャッシュ0, 1 (612-0, 1) の配下にそれぞれ100KBのホストデータが格納される。その後の処理は前記1.8MBの転送例と同様である。

【0038】次に、リードヒット/ミス判定はSGCB0のLBAとセグメント情報より行われる。ホストからのリードコマンドがデータ配分制御部から、上位ファイバインタフェース0に送られる。要求LBAと要求データがキャッシュ上に存在する場合は、ヒットと判定し即座にデータ転送を始めることが可能である。要求LBAが存在するが、要求データの一部が既にHDDに書きこまれて、そのデータが使用していたセグメントを他のデータに割り当てて、キャッシュ上にデータの一部を存在しない場合は、キャッシュヒットと判定するが、キャッシュ上に欠ける分のデータをHDDよりリードしてからデータ転送を始める。ホストコマンドが要求するLBAとSGCB0が持っているLBA情報と一致しない場合はミスと判定する。

【0039】ホストの要求により、前記に保存した1.8MBデータをディスクアレイより読み出す時の流れを述べる。

【0040】データ配分制御部でホストからのリードコマンドを解釈し、読み出し先、要求データ長を割り出して一時保存する。要求読み出しデータ長を上位ファイバチャンネルの数で割ったものを各上位ファイバドライブ制御部が転送するデータ長とする。そして、上位ファイバインタフェース0に読出先のアドレスLBA、転送データ長情報を含むリードコマンドを発行する。キャッシュ0で前記規則でヒット／ミスの判定をし、ヒットの場合はキャッシュ0、1からデータを返す。

【0041】ミスの場合はリードアドレスLBA、データ転送長情報をHDDインタフェース106に送り、106より、ホストデータの半分、ブロック0をキャッシュ0に送る。その後、キャッシュ0からキャッシュ1経由でHDDインタフェース106にホストデータの半分、ブロック1をリードするアドレスLBA、伝送データ長情報を含むリードコマンドを発行する。ブロック0の転送完了後、106よりブロック1をキャッシュ1へ転送する。ブロック1のリードLBAはブロック0のLBA+転送データ長とする。

【0042】キャッシュにリードデータを入れた後、データ配分制御部にデータを転送するために、前記ファイバドライブ制御部0、1に、キャッシュからのデータをバッファに入れる順番に関する図9の転送リストを生成する。データ配分制御部バッファに入れたキャッシュからのデータをホストに返すためにホストファイバドライブ制御部301にも図10に示す形の転送リストを生成する。

【0043】キャッシュ側のファイバドライブ制御部0'、1'（610-0、1）がデータ配分制御部からのリードコマンドを受けとり、HDDインタフェース0、1にリード先アドレス、リードデータ長を指示することにより、前記HDDインタフェース0、1はキャッシュ0にブロック0を、キャッシュ1にブロック1を転送する。その後、ファイバドライブ0'、1'より、データブロック0、1をデータ配分制御部へ送信する。上位ファイバドライブ0、1制御部（307-0、1）が受け取ったデータを図9の転送リストにしたがって3回ずつデータ配分制御のバッファへデータ転送を行う。

【0044】上位ファイバドライブ制御部0、1による一回目の転送では、バッファ0、1（304-0、1）が満杯になった後、上位ファイバドライブ制御部0、1が前記バッファ0、1を放し、バッファ使用状態のフラグが1にしてから、ホストファイバドライブ301が起動され、バッファ0、1の順からホストにデータを返す。その同時、上位ファイバドライブ0、1よりバッファ2、3へ2回目のデータ転送を行い、完了後、該当バッファのフラグビットを1にしてホストファイバドライブ制御部が使用可能な状態にしておく。バッファ0、1からホストへのデータ転送が終わり、使用可能状態になると、再びバッファ0、1に、上位ファイバドライブ制

御部0、1より最後の600MBを格納される。それと同時に、ホストファイバドライブ制御部301がバッファ2、3よりホストへデータを返し、最後の600MBをバッファ0、1よりホストへ転送する。

【0045】続いて、前記例で保存した200MBデータをHDD群より読み出し時の流れを述べる。データ配分制御部203でホストからのリードコマンドを解釈し、読み出し先、要求データ長を割り出して一時保存する。要求読み出しデータ長を上位ファイバインタフェース（608-0、1）の数で割ったもの、100MBを上位ファイバドライブが転送するデータ長とし、上位ファイバドライブ制御部0、1の制御部に、バッファ0、1のみ一回ずつ使用する。データ配分制御部は上位ファイバインタフェース0に読出先のアドレスLBA、転送データ長情報を含むリードコマンドを発行する。

【0046】キャッシュ0で前記規則でヒット／ミスの判定をし、ヒットの場合はキャッシュからデータを返す。ミスの場合はリードアドレスLBA、データ転送長情報をHDDインタフェース106に送り、106より、HDDから100MBのデータブロック0をキャッシュ0、ブロック1をキャッシュ1に格納する。キャッシュ0、1から各バッファにおいて100MBのデータを送信のための転送リストが生成される。データ配分制御部バッファにキャッシュからのデータをホストに返すためにホストファイバ制御部と上位ファイバインタフェース制御部に、図9、図10と同様な転送リストを生成する。

【0047】前記転送リストにしたがって、上位ファイバドライブ制御部0、1はデータ配分制御のバッファ0、1へ同時に100MBのデータを転送する。バッファ0、1がホストファイバドライブ制御部が使用できる状態になった後はバッファ0、1から100MBずつのデータをホストに返す。

【0048】本実施形態にあるHDD群はファイバインタフェースを有するHDD採用しており、HDDインタフェースにHDD群を制御するためのファイバドライブFPCが使用される。パリティ生成機能をキャッシュコントローラに持たした形となっているが、キャッシュコントローラではなく、HDDインタフェースにパリティ生成機能を持たすことも可能である。

【0049】また、本実施形態において、ディスクアレイ内部に二つの上位ファイバインタフェースを使用する例を挙げているが、上位ファイバインタフェースの数は二つに限らず、複数であればよい。また、本実施形態において、データ配分制御部とキャッシュ間はファイバチャンネルを介して接続されるが、直接内部バスで接続される場合も、同様な効果は得られる。

【0050】以上説明したように、本発明の実施形態は、次のような構成、機能乃至作用を奏するものを含むものである。即ち、ホストファイバインタフェースから

のデータを授受する複数の上位ファイバインタフェースと、前記複数上位ファイバインタフェースに接続されたキャッシュメモリと、データを書き込み又は読み出す複数のディスク記憶手段と、前記記憶手段を制御するディスクドライブインタフェースと、を備えたディスクアレイシステムにおいて、前記ホストファイバインタフェースと前記上位ファイバインタフェースとの間に設けられたデータ配分制御部は、ホストファイバインタフェースからのデータを上位インタフェースの数に応じて複数ブロックに分割し、データ配分制御部に接続される複数の上位ファイバインタフェースを介して前記分割されたデータを並列に複数からなるキャッシュメモリに同時に転送を行うことと、前記ディスクアレイ内部が複数ファイバインタフェースがそれぞれのキャッシュメモリを持つことにより、上位ファイバインタフェースに繋ぐディスクアレイ内部バスの転送速度（即ち、図 1 に示す各内部バス 109-0 又は 109-1 の転送速度）がホストファイバインタフェースに比べてその転送速度が低くても転送速度が高速であるホストファイバインタフェースへの対応を可能とすることである。

【0051】また、本実施形態は、データ配分制御部は自身が持つファイバインタフェースの ID をホストに見せ、ホストからのアクセス要求を一旦、バッファで受け取り分割し、分割したアクセス要求を同時に上位インタフェースに発行して実行させることにより、ホストがディスクアレイ中の複数ファイバインタフェースの存在を意識させないこととなる。換言すると、ファイバループ又はファイバスイッチによる接続と異なり、ホストファイバインタフェースがバッファを噛まして複数の上位ファイバチャンネルと接続し、ホストファイバインタフェースは上位ファイバインタフェースと直接に接続されないこととなる。

【0052】

【発明の効果】本願において開示される発明の内、体系的なものによって得られる効果を簡単に説明すれば、下記の通りである。

【0053】ディスクアレイ内部の複数の上位ファイバインタフェースを同時に動作させることより、ホストファイバインタフェースに比べてその転送速度の遅い内部バスを有するディスクアレイシステムでも、転送速度が高速であるホストファイバインタフェースを有するホストからの高速アクセスに速度を損うことなく、実行することが可能となった。

【0054】また、データ配分制御部のバッファにおいて、ホストライト要求にキャッシュの空き容量で答えることと、複数面の小容量なバッファを使いまわしすることにより、小容量なバッファでホストとディスクアレイ間との大容量なデータ転送を可能にした。

【0055】また、多面キャッシュを一元管理することで、現用キャッシュ制御マイクロに大きな変更を加えず

に流用することが可能である。

【図面の簡単な説明】

【図 1】本発明の実施形態に係るディスクアレイシステムの構成の概要を示す概念図である。

【図 2】RAID 技術における従来のディスクアレイシステムの概略図である。

【図 3】本実施形態のデータ配分制御部の構成を示す図である。

【図 4】本実施形態におけるライト時のデータ配分制御部でのバッファ動作のタイミングを示す図である。

【図 5】本実施形態におけるライト時のデータ配分制御部動作と、ホスト又は内部上位ファイバインタフェース間動作とのタイムチャート図である。

【図 6】本実施形態におけるデータ配分制御部とキャッシュ間の構成を示す図である。

【図 7】本実施形態におけるキャッシュ管理テーブル SGCB0、1 及びセグメントのキャッシュへの格納イメージ図である。

【図 8】本実施形態におけるキャッシュライト時の上位ファイバドライブ転送用転送リストを示す図である。

【図 9】本実施形態におけるキャッシュリード時の上位ファイバドライブ転送用転送リストを示す図である。

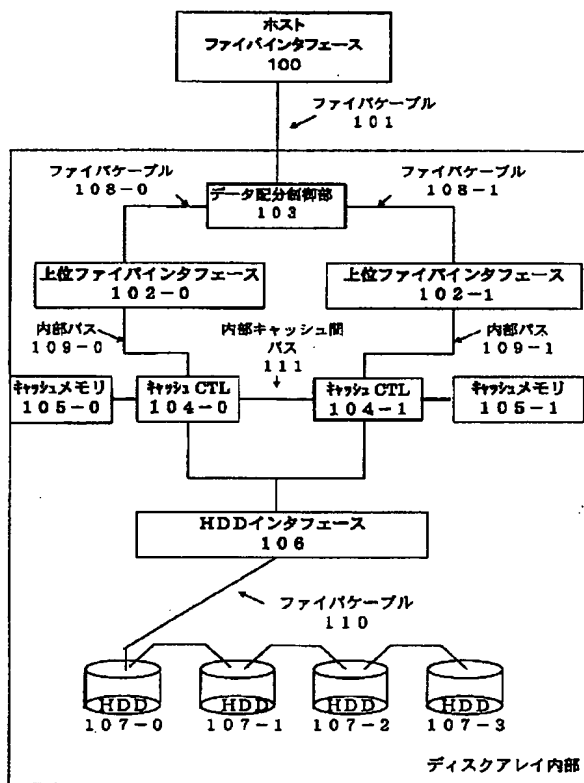
【図 10】本実施形態におけるリード時のホストファイバドライブ転送用転送リストを示す図である。

【符号の説明】

- 100 ホストファイバインタフェース
- 101 ファイバケーブル
- 102-0~1 上位ファイバインタフェース
- 103 データ配分制御部
- 104-0~1 キャッシュコントローラ
- 105-0~1 キャッシュメモリ
- 106 HDDインタフェース
- 107-0~3 HDDドライブ
- 108-0~1 ファイバケーブル
- 109-0~1 内部バス
- 110 ファイバケーブル
- 111 内部キャッシュ間バス
- 200 ホストインタフェース
- 201 Fibre/SCSI ケーブル
- 202 上位インタフェース
- 203 内部バス
- 204 MPU ユニット
- 205 キャッシュコントローラ
- 206 キャッシュメモリ
- 207 HDDインタフェース
- 208 Fibre/SCSI ケーブル
- 209-0~3 HDDドライブ
- 301 ホストファイバドライブ制御部
- 302 バッファバススイッチ
- 303 内部バス

304-0~3 データバッファ
 305-0~3 内部バス
 306 バッファバススイッチ
 307-0~1 上位ファイバドライブ制御部
 601 ホストファイバドライブ制御部
 602 バッファバススイッチ
 603 内部バス
 604-0~3 データバッファ
 605 バッファバススイッチ

【図1】

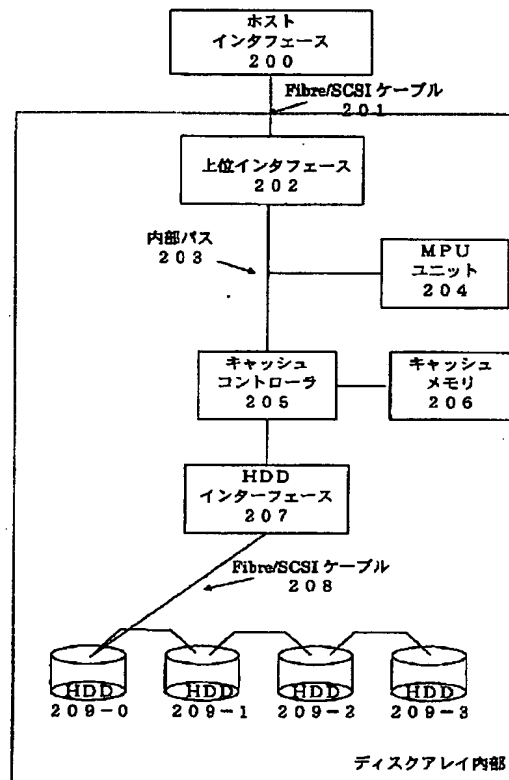


【図9】

| 上位ファイバドライブ制御部転送リスト | | | | |
|--------------------|------------|-------|-----------------------------|---------|
| 転送リスト | 送信ファイバドライブ | 送信先 | 読み出し先 | 転送データ長 |
| 1 | 0 | バッファ0 | ブロック0 : SEGR _{xxx} | 計300 MB |
| | 1 | バッファ1 | ブロック1 : SEGL _{xxx} | 計300 MB |
| 2 | 0 | バッファ2 | ブロック0 : SEGR _{xxx} | 計300 MB |
| | 1 | バッファ3 | ブロック1 : SEGL _{xxx} | 計300 MB |
| 3 | 0 | バッファ0 | ブロック0 : SEGR _{xxx} | 計300 MB |
| | 1 | バッファ1 | ブロック1 : SEGL _{xxx} | 計300 MB |

606~7 内部バス
 608-0~1 上位ファイバドライブ0, 1制御部
 609 MPUユニット部
 610-0~1 上位ファイバドライブ0', 1' 制御部
 611-0~1 ファイバケーブル
 612-0~1 キャッシュコントローラ及びキャッシュメモリ
 613-0~1 内部バス

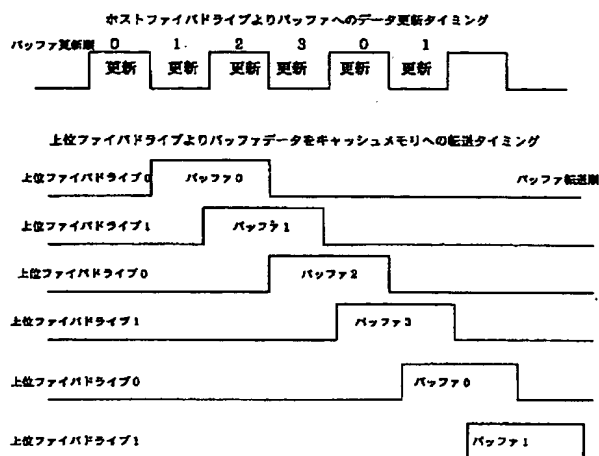
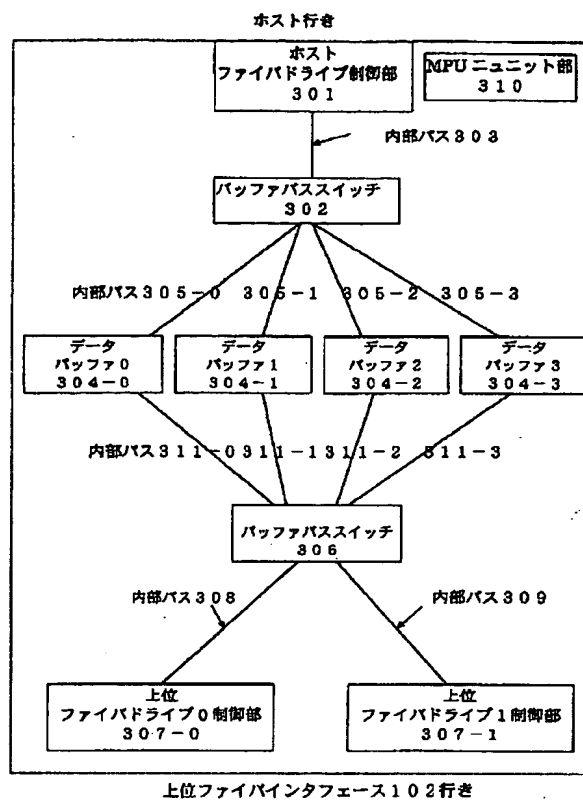
【図2】



【図10】

| 高速ホストファイバドライブ制御部転送リスト | | |
|-----------------------|-------|----------|
| ホストファイバドライブ転送リスト | 読み出し先 | 読み出しデータ長 |
| 0 | バッファ0 | 全領域 |
| 1 | バッファ1 | 全領域 |
| 2 | バッファ2 | 全領域 |
| 3 | バッファ3 | 全領域 |
| 4 | バッファ0 | 全領域 |
| 5 | バッファ1 | 全領域 |

【图 4】



【図 8】

(1)

ファイバドライブ 307-0 における制御部転送リスト

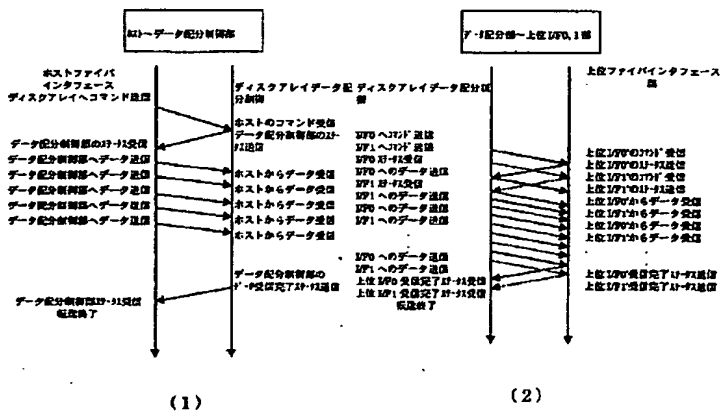
| 転送リスト | 307-0 | 読み出し先 | 転送先 |
|-------|-------|----------------|--------|
| 1 | 起動 | バッファ0 304-0 | キャッシュ0 |
| 2 | 起動 | バッファ2 304-2 | キャッシュ0 |

【図 5】

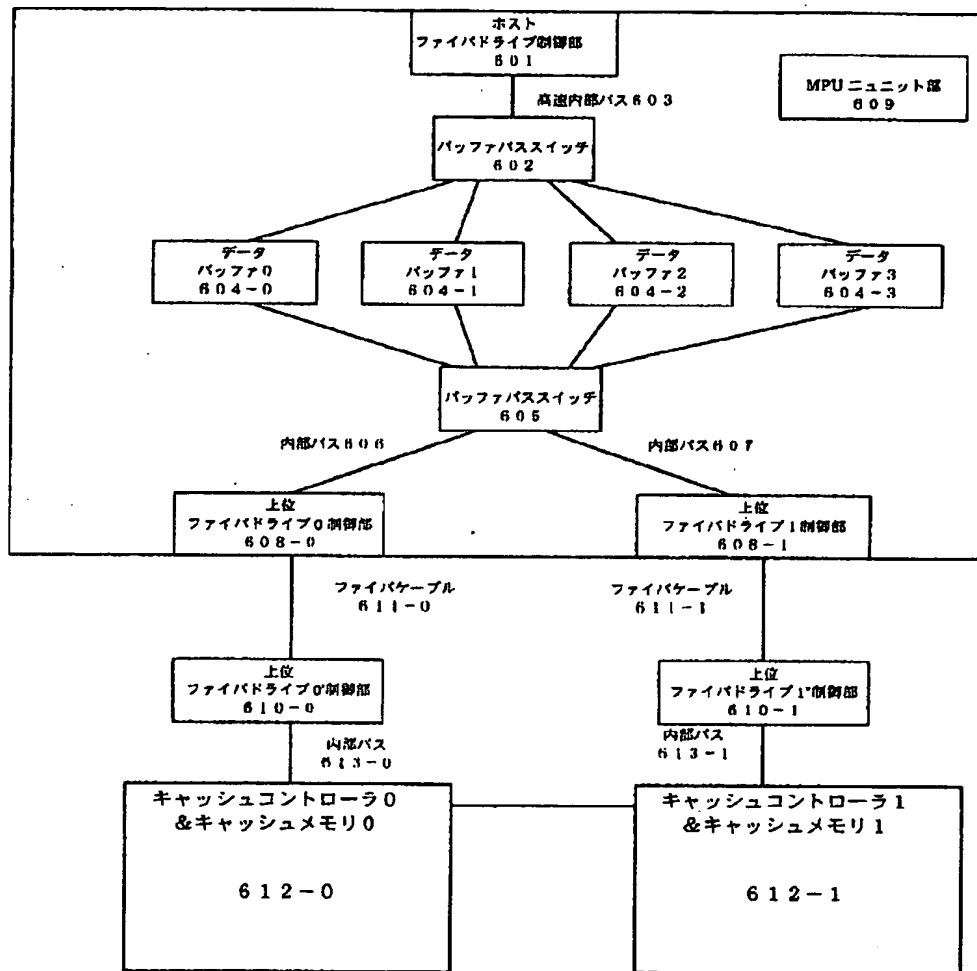
(2)

ファイバドライブ307-1における制御部転送リスト

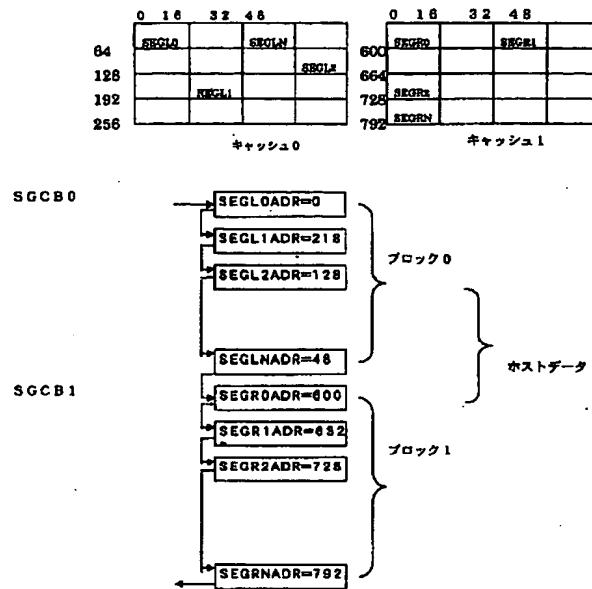
| 転送リスト | 307-1 | 読み出し先 | 転送先 |
|-------|-------|----------------|--------|
| 1 | 起動 | バッファ1 304-1 | キャッシュ1 |
| 2 | 起動 | バッファ4 304-3 | キャッシュ1 |



【図6】



【図7】



フロントページの続き

(72)発明者 八木沢 育哉

神奈川県川崎市麻生区王禅寺1099番地 株
 式会社日立製作所システム開発研究所内

Fターム(参考) 5B005 JJ12 MM12 NN12

5B065 BA01 CA30 CC08 CE12 CE14
 CE26